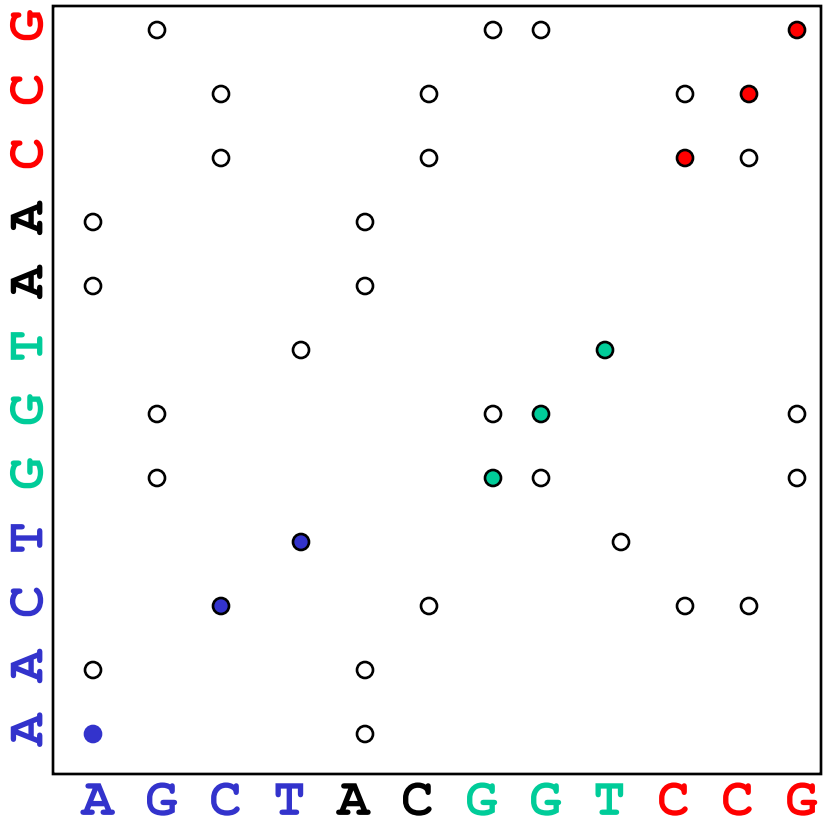
Recherche par similarité dans les banques/bases de données
La suite Blast (Basic Local Alignment Search Tool)


G
C
C
A
A
T
G
G
T
C
C
A
A
C
T
G
G
T
C
C
G
A
G
C
T
A
C
G
G
T
C
C
G

Algorithme de Blast (version 1)

(Altschul *et al.* (1990) J. Mol. Biol., 215, 403-410)

- ne recherche que les diagonales

-> établir un score pour chaque paire de segments comparée (diagonale)

(séquences ADN : Se(identité) = 5 et Se(mismatch) = -4

séquences protéiques -> matrice de substitution)

-> MSP : Maximal Segment Pair correspond à la paire de segments (diagonale) qui possède le score le plus élevé comparé au score de toutes les autres paires de segments.

-> HSP : High Scoring segment Pair définit une paire de segments comme localement maximale, c'est-à-dire que l'on ne pourra pas augmenter son score soit en la rallongeant, soit en les raccourcissant.

-> problème : trop de paires de segments, on ne peut pas toutes les retenir.

-> définition d'un seuil S_{HSP} : Blast ne retient comme HSP que les paires de segments dont le score est supérieur à S_{HSP}.

Algorithme de Blast (version 1)

(Altschul *et al.* (1990) J. Mol. Biol., 215, 403-410)

-> Définition de la valeur du seuil S_{HSP} :

Résultats en statistique de Karling *et al.* permettent d'estimer le score le plus élevé que peut avoir une MSP entre deux séquences dont la similarité ne serait due qu'au hasard

-> ce score est utilisé comme seuil S_{HSP}

-> Sélection des HSP

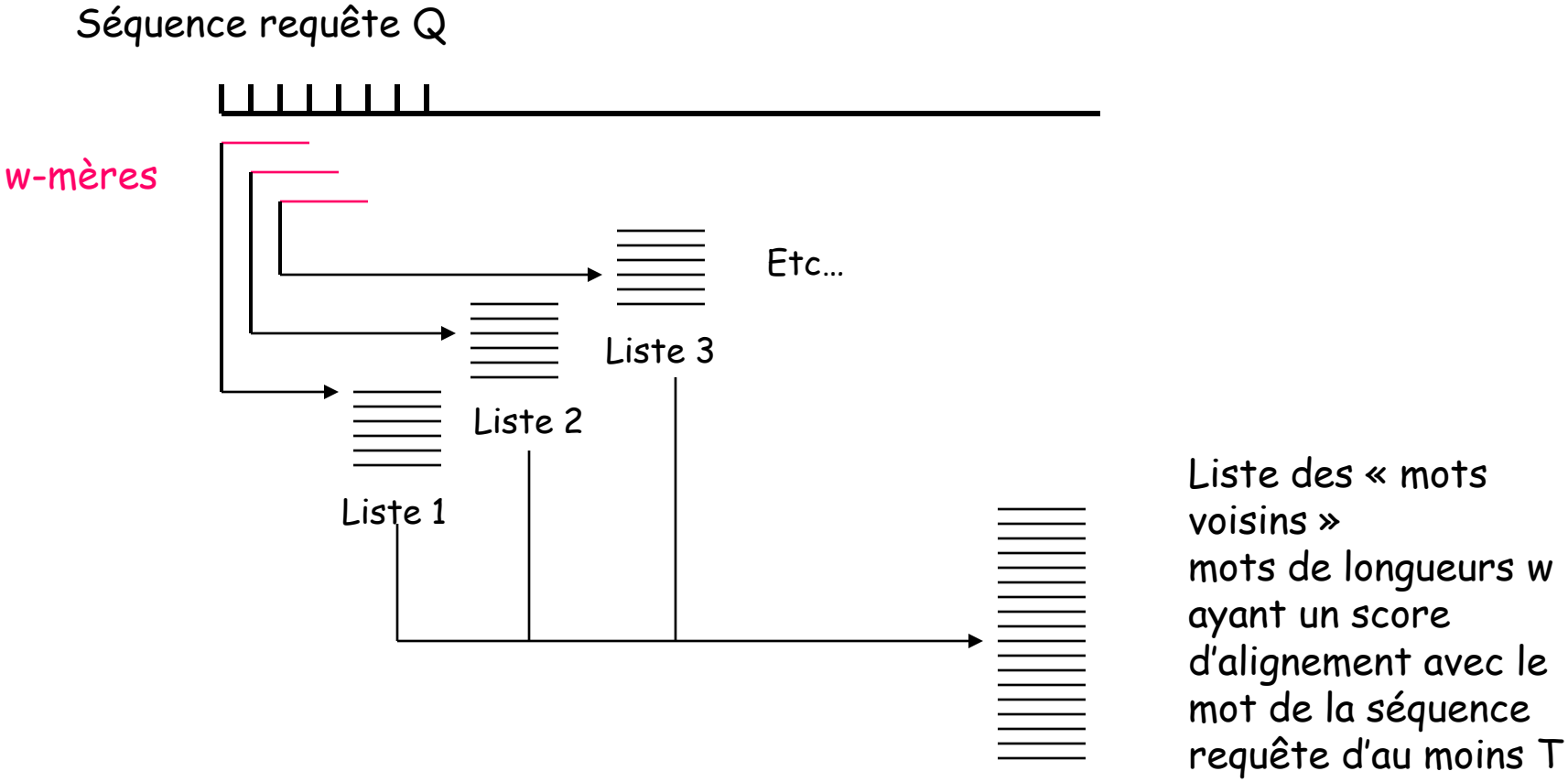
• Paire de segments dont le score est < à S_{HSP}, similarité non significative due au hasard. Cette paire n'est pas retenue.

• Paire de segments dont le score est $\geq$ à S_{HSP}, similarité due à une histoire évolutive commune -> séquence apparentée -> retenue comme HSP

Algorithme de Blast (version 1)

(Altschul *et al.* (1990) J. Mol. Biol., 215, 403-410)

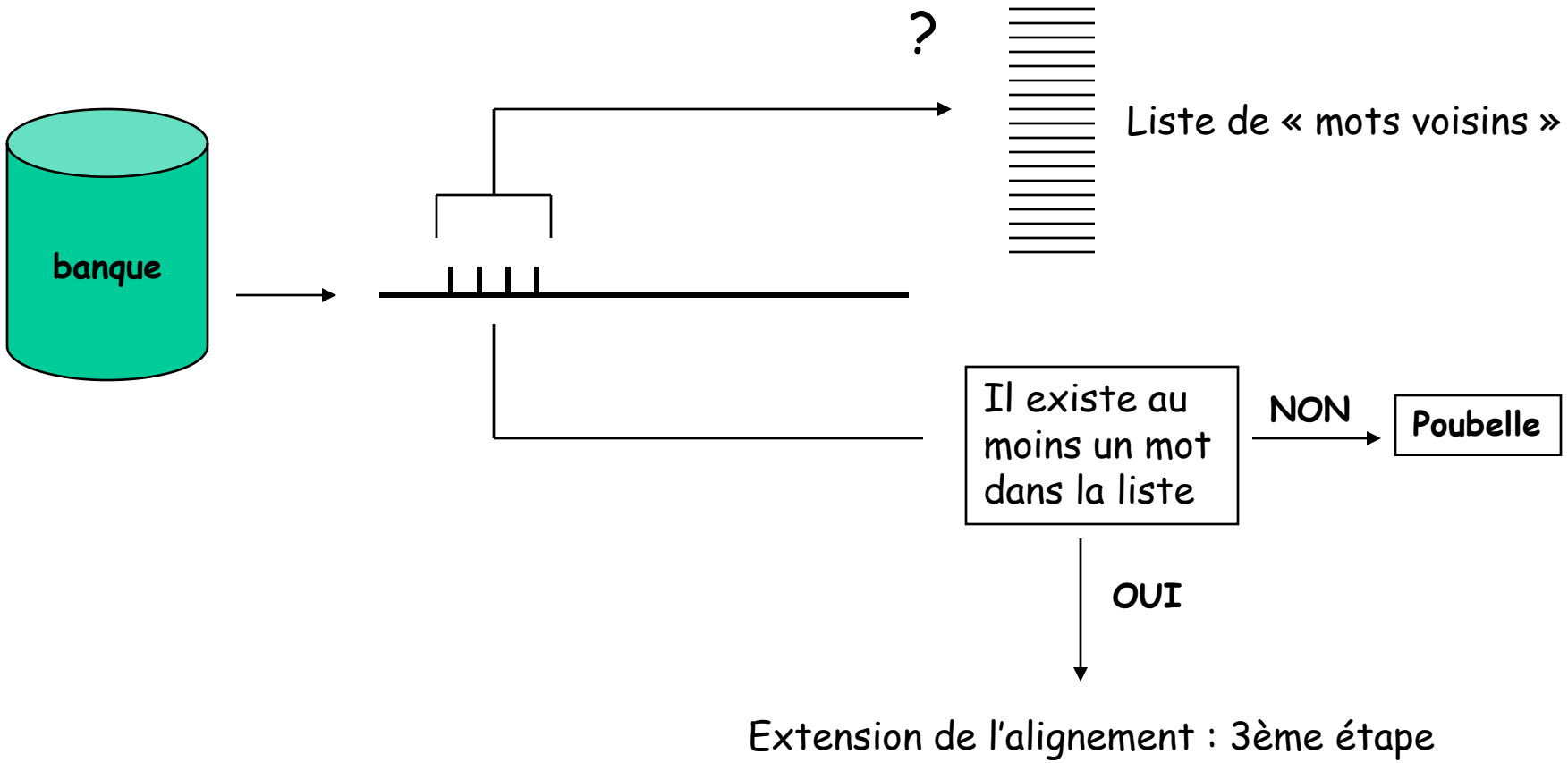
1ère étape : Établissement d'une liste de « mots voisins »



w-mères : mot de taille w (w = 3 pour les protéines et w = 11 pour les séquences d'acides nucléiques)

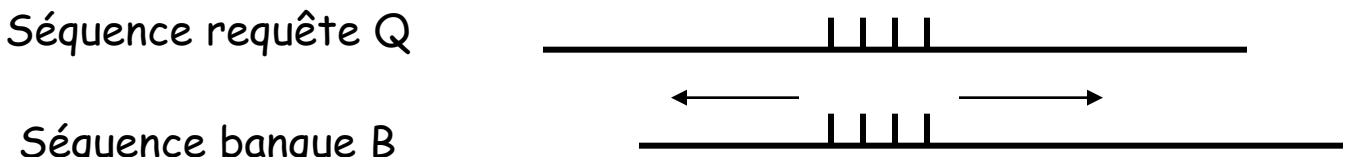
Algorithme de Blast (version 1)

2ème étape : Recherche de la présence de ces mots dans les séquences de la banque



Algorithme de Blast (version 1)

3ème étape : si un « hit » extension de l'alignement



Essayer d'étendre l'alignement dans les deux sens (arrêt de l'extension quand le score obtenu décroît au minimum d'une valeur X fixée (drop-off score)

Il existe au moins un prolongement d'un score > S

NON → **Poubelle**

OUI

Meilleur score obtenu par prolongement d'un hit → Score séquence B de la banque

Statistique Poisson

4ème étape : Calcul de la p-value →

Classement des séquences de la banque

Normalisation du score

Le score normalisé S' pour un HSP est donné par :

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

Avec :
S score du HSP
λ et K paramètres calculés à partir du système de score et de la composition des séquences

S' est exprimé en une unité appelée *bits*

La p-value

La p-value est la probabilité qu'il existe au moins un HSP, obtenu lors de la comparaison de deux séquences aléatoires (de même longueur et composition que les séquences d'intérêt), dont le score soit supérieur ou égal à celui du HSP issu de la comparaison des séquences d'intérêt.

La p-value

On compare une séquence de longueur n avec une banque de données de longueur m, avec une matrice de comparaison donnée.

On s'intéresse à la distribution du nombre $N(S)$ d'alignements dont le score dépasse la valeur S. Cette distribution suit approximativement une loi de Poisson. On a donc l'expression suivante pour la probabilité qu'il existe un alignement fortuit de score $\geq S$:

$$p(score \geq S) = 1 - e^{-E(S)}$$

$E(S)$ étant l'espérance mathématique de $N(S)$ et qui s'exprime par :

$$E(S) = Knme^{-\lambda S}$$ = E-value = nombre attendu d'alignements ayant un score au moins égal à S

Avec λ et K paramètres calculés à partir du système de score et de la composition des séquences

Dans le cas des alignements locaux sans gap, la théorie de Karlin et Altschul fournit donc des formules analytiques dont les paramètres dépendent de ceux de la recherche.

Résultat d'une recherche avec la version 1 de BlastP

```
Query= Bsub.RbsA
         (493 letters)

Database: Coli
           4289 sequences; 1,358,990 total letters

Searching.................................................done

                                                                 Score     E
Sequences producing significant alignments:                      (bits)  Value  N

EcolA01.RBSA "ATP-binding component of D-ribose high-affini...   270  e-141  3
EcolA01.MGLA "ATP-binding component of methyl-galactoside t...   164  e-124  4
EcolA01.YJCW "putative ATP-binding component of a transport...   145  e-121  4
EcolA01.XYLG "putative ATP-binding protein of xylose transp...   157  e-120  6

>EcolA01.RBSA "ATP-binding component of D-ribose high-affinity
            transport system"
           Length = 501

 Score =  270 bits (587), Expect(3) = e-141
 Identities = 115/258 (44%), Positives = 167/258 (64%)

Query: 3   IEMKDIHKTFGKNQVLSGVSFQLMPGEVHALMGENGAGKSTLMNILTGLHKADKGQISIN 62
           +++K I K F   + LSG +  + PG V AL+GENGAGKST+M +LTG++  D G +
Sbjct: 5   LQLKGIDKAFPGVKALSGAALNVYPGRVMALVGENGAGKSTMMKVLTGIYTRDAGTLLWL 64

Query: 63  GNETYFSNPKEAEQHGIAFIHQELNIWPEMTVLENLFIGKEISSKLGVLQTRKMKALAKE 122
           G ET F+ PK +++ GI  IHQELN+ P++T+ EN+F+G+E  ++ G +  + M A A +
Sbjct: 65  GKETTFTGPKSSQEAGIGIIHQELNLIPQLTIAENIFLGREFVNRFGKIDWKTMYAEADK 124

Query: 123 QFDKLSVSLSLDQEAGECSVGQQQMIEIAKALMTNAEVIIMDEPTAALTEREISKLFEVI 182
              KL++    D+  G+ S+G QQM+EIAK L   ++VIIMDEPT ALT+ E   LF VI
Sbjct: 125 LLAKLNLRFKSDKLVGDLSIGDQQMVEIAKVLSFESKVIIMDEPTDALTDTETESLFRVI 184

Query: 183 TALKKNGVSIVYISHRMEEIFAICDRITIMRDGKTVDTTNISETDFDEVVKKMVGRELTE 242
             LK  G  IVYISHRM+EIF ICD +T+ RDG+ +    ++    D +++ MVGR+L +
Sbjct: 185 RELKSQGRGIVYISHRMKEIFEICDDVTVFRDGQFIAEREVASLTEDSLIEMMVGRKLED 244

Query: 243 RYPKRTPSLGDKVFEVKN 260
           +YP    + GD   +V N
Sbjct: 245 QYPHLDKAPGDIRLKVDN 262


 Score =  103 bits (220), Expect(3) = e-141
 Identities = 43/94 (45%), Positives = 66/94 (69%)

Query: 269 DVSFYVRSGEIVGVSGLMGAGRTEMMRALFGVDRLDTGEIWIAGKKTAIKNPQEAVKKGL 328
           DVSF +R GEI+GVSGLMGAGRTE+M+ L+G     +G + + G +   ++PQ+ +  G+
Sbjct: 270 DVSFTLRKGEILGVSGLMGAGRTELMKVLYGALPRTSGYVTLDGHEVVTRSPQDGLANGI 329

Query: 329 GFITENRKDEGLLLDTSIRENIALPNLSSFSPKG 362
           G I+E+RK +GL+L   S++EN++L   L  FS  G
Sbjct: 330 VYISEDRKRDGLVLGMSVKENMSLTALRYFSRAG 363
```

```
 Score =  165 bits (356), Expect(3) = e-141
 Identities = 74/130 (56%), Positives = 92/130 (69%)

Query: 362 GLIDHKREAEFVDLLIKRLTIKTASPETHARHLSGGNQQKVVIAKWIGIGPKVLILDEPT 421
           G + H  E + V   I+   +KT S E     LSGGNQQKV IA+ +   PKVLILDEPT
Sbjct: 364 GSLKHADEQQAVSDFIRLFNVKTPSMEQAIGLLSGGNQQKVAIARGLMTRPKVLILDEPT 423

Query: 422 RGVDVGAKREIYTLMNELTERGVAIIMVSSELPEILGMSDRIIVVHEGRISGEIHAREAT 481
           RGVDVGAK+EIY L+N+    G++II+VSSE+PE+LGMSDRIIV+HEG +SGE    +AT
Sbjct: 424 RGVDVGAKKEIYQLINQFKADGLSIILVSSEMPEVLGMSDRIIVMHEGHLSGEFTREQAT 483

Query: 482 QERIMTLATG 491
           QE +M  A G
Sbjct: 484 QEVLMAAAVG 493
```

Gapped Blast ou Blast2

(Altschul *et al.* (1997) Nucleic Acids Res., 25, 3389-3402)

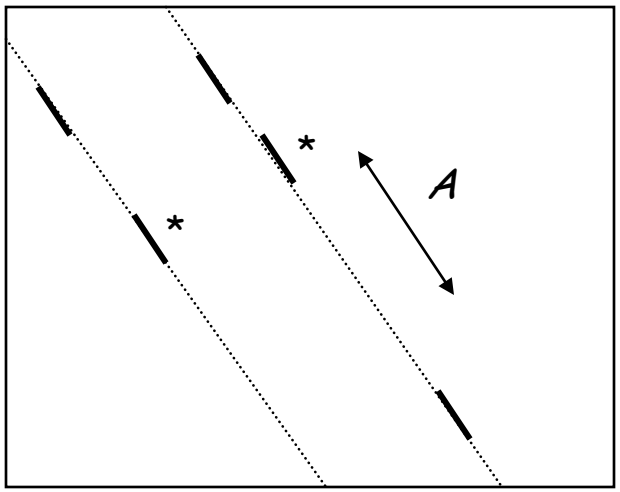
Les deux premières étapes identiques à la version 1 de l'algorithme de Blast

Différences :

- étape d'extension des hits (90% du temps d'exécution du Blast)

Rq : une HSP contient souvent plusieurs hits sur la même diagonale et à une distance peu éloignée les uns des autres.

Sélection des hits pour l'étape d'extension : présence de deux hits sur la même diagonale séparés par une distance inférieure ou égale à *A*.



Les hits marqués d'une * subiront une extension sans gaps analogue à la 3ème étape du Blast version 1

Gapped Blast ou Blast2

Différences :

• Les HSP sélectionnés qui auront un score supérieur ou égal à S servent ensuite de points d'ancrage à une recherche d'alignement local optimal par programmation dynamique. La matrice de programmation dynamique est explorée dans les deux directions à partir d'un résidu aligné (graine).
Choix de la graine : on choisit le long d'une HSP le segment de 11 résidus ayant le meilleur score et le résidu central est utilisé comme graine.
La recherche du chemin optimal est limité aux cellules de la matrice telles que le score de l'alignement ne devienne pas inférieur de plus de Xg au score maximal atteint jusque là (modification de l'algorithme de Smith et Waterman)

Significativité des alignements : La E-value

Dans le cas des alignements locaux avec gaps, il n'existe pas de théorie décrivant la distribution attendue des scores.

La E-value (Expect value), le nombre attendu d'alignement qui par chance aurait un score $\geq S$, est évaluée en regardant les scores des alignements générés par comparaison de séquences aléatoires ayant même longueur et même composition que la séquence requête.

La suite Blast

Un ensemble de programmes :

programme	séquence requête	Banque
BlastN	nucléique	nucléique
BlastP	protéique	protéique
BlastX	nucléique (séquence traduite dans les 6 cadres de lecture)	protéique
tblastN	protéique	nucléique (séquences de la banque traduites dans les 6 cadres de lecture)
tblastX	nucléique (séquence traduite dans les 6 cadres de lecture)	nucléique (séquences de la banque traduites dans les 6 cadres de lecture)

La suite Blast

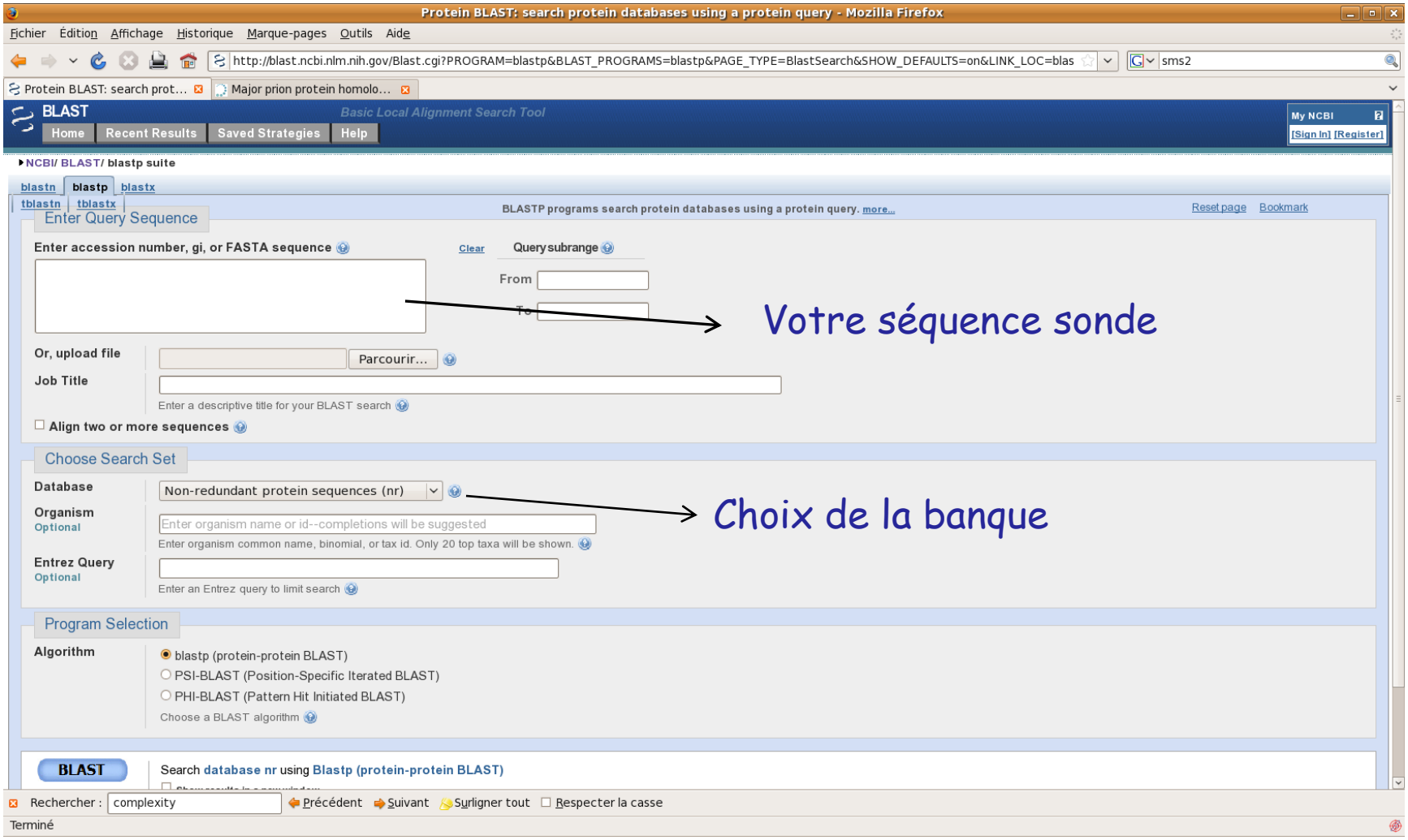
Possibilité d'appliquer des filtres et masques (paramètres de l'algorithme) :

- masquer les séquences de faible complexité (proposé pour l'ensemble des programmes de la suite Blast)

- dans le cas d'une recherche avec une séquence d'acides nucléiques contre une banque de séquences nucléiques (BlastN), masquer les séquences répétées (ex: les séquences Alu chez les primates).

  Une première analyse compare la séquence sonde à une banque de séquences d'éléments répétés. Les zones de la séquence sonde s'alignant avec les séquences d'éléments répétés sont masquées pour la recherche de similarité dans la banque.

Exemple d'interface du programme BlastP (site NCBI)



Les paramètres « cachés »

Nombre de séquences cibles

Seuil E-value

Taille des mots pour construire la liste des mots voisins

Choix de la matrice de substitution

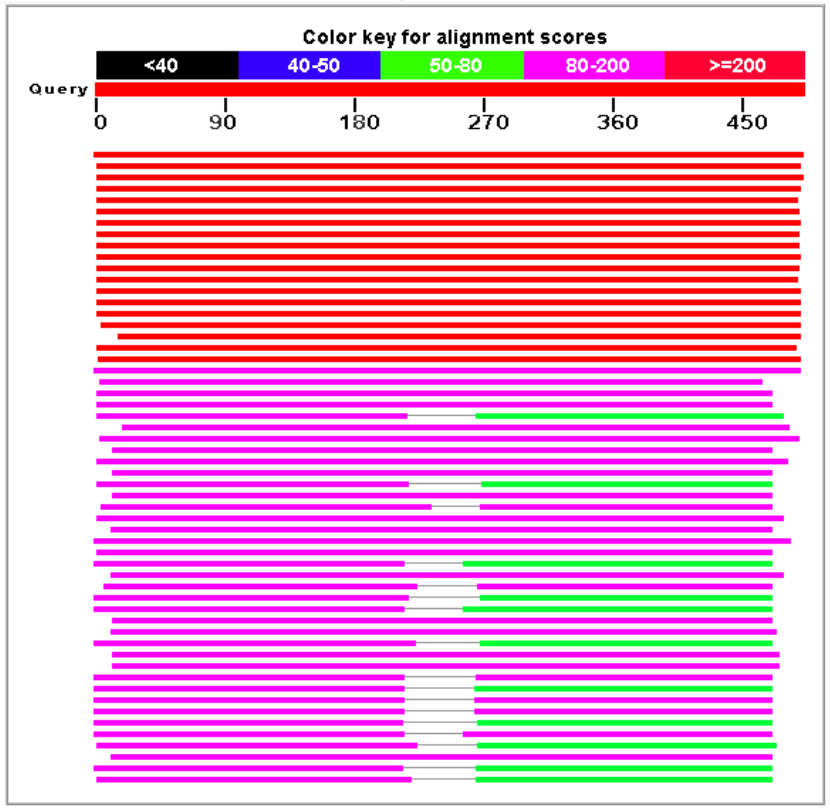
Pondération des gaps : ouverture et extension

Résultat d'une recherche avec BlastP2 sur le site du NCBI

Query= Bsub.RbsA
Length=493

Distribution of 205 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Distance tree of results NEW

Sequences producing significant alignments:		Score (Bits)	E Value
gi|7404442|sp|P36947|RBSA_BACSU	Ribose transport ATP-binding pro	932	0.0
gi|132119|sp|P04983|RBSA_ECOLI	Ribose transport ATP-binding prot	432	2e-120
gi|1172865|sp|P44735|RBSA_HAEIN	Ribose transport ATP-binding pro	416	1e-115
gi|6016558|sp|Q56342|MGLA_TREPA	Galactoside transport ATP-bindin	400	8e-111
gi|731981|sp|P32721|ALSA_ECOLI	D-allose transport ATP-binding pr	395	2e-109
gi|2506104|sp|P23924|MGLA_SALTY	Galactoside transport ATP-bindin	395	2e-109
gi|1170944|sp|P44884|MGLA_HAEIN	Galactoside transport ATP-bindin	391	3e-108
gi|20137563|sp|Q9S472|ARAG_BACST	L-arabinose transport ATP-bindi	390	5e-108
gi|77416578|sp|P0AAG8|MGLA_ECOLI	Galactoside transport ATP-bi...	390	6e-108

Résultat d'une recherche avec BlastP2 sur le site du NCBI (suite)

```
>gi|132119|sp|P04983|RBSA_ECOLI  Ribose transport ATP-binding protein rbsA
Length=501

 Score =  432 bits (1110),  Expect = 2e-120, Method: Composition-based stats.
 Identities = 232/490 (47%), Positives = 325/490 (66%), Gaps = 2/490 (0%)

Query  3    IEMKDIHKTFGKNQVLSGVSFQLMPGEVHALMGENGAGKSTLMNILTGLHKADKGQISIN  62
            +++K I K F   + LSG +  + PG V AL+GENGAGKST+M +LTG++  D G + +
Sbjct  5    LQLKGIDKAFPGVKALSGAALNVYPGRVMALVGENGAGKSTMMKVLTGIYTRDAGTLLWL  64

Query  63   GNETYFSNPKEAEQHGIAFIHQELNIWPEMTVLENLFIGKEISSKLGVLQTRKMKALAKE  122
            G ET F+ PK +++ GI  IHQELN+ P++T+ EN+F+G+E  ++ G +  + M A A +
Sbjct  65   GKETTFTGPKSSQEAGIGIIHQELNLIPQLTIAENIFLGREFVNRFGKIDWKTMYAEADK  124

Query  123  QFDKLSVSLSLDQEAGECSVGQQQMIEIAKALMTNAEVIIMDEPTAALTEREISKLFEVI  182
               KL++    D+  G+ S+G QQM+EIAK L   ++VIIMDEPT ALT+ E   LF VI
Sbjct  125  LLAKLNLRFKSDKLVGDLSIGDQQMVEIAKVLSFESKVIIMDEPTDALTDTETESLFRVI  184

Query  183  TALKKNGVSIVYISHRMEEIFAICDRITIMRDGKTVDTTNISETDFDEVVKKMVGRELTE  242
              LK  G  IVYISHRM+EIF ICD +T+ RDG+ +    ++    D +++ MVGR+L +
Sbjct  185  RELKSQGRGIVYISHRMKEIFEICDDVTVFRDGQFIAEREVASLTEDSLIEMMVGRKLED  244

Query  243  RYPKRTPSLGDKVFEVKNASVKGSFEDVSFYVRSGEIVGVSGLMGAGRTEMMRALFGVDR  302
            +YP    + GD   +V N    G   DVSF +R GEI+GVSGLMGAGRTE+M+ L+G
Sbjct  245  QYPHLDKAPGDIRLKVDNLCGPG-VNDVSFTLRKGEILGVSGLMGAGRTELMKVLYGALP  303

Query  303  LDTGEIWIAGKKTAIKNPQEAVKKGLGFITENRKDEGLLLDTSIRENIALPNLSSFS-PK  361
              +G + + G +   ++PQ+ +  G+ +I+E+RK +GL+L  S++EN++L  L  FS
Sbjct  304  RTSGYVTLDGHEVVTRSPQDGLANGIVYISEDRKRDGLVLGMSVKENMSLTALRYFSRAG  363

Query  362  GLIDHKREAEFVDLLIKRLTIKTASPETHARHLSGGNQQKVVIAKWIGIGPKVLILDEPT  421
            G + H  E + V   I+   +KT S E     LSGGNQQKV IA+ +   PKVLILDEPT
Sbjct  364  GSLKHADEQQAVSDFIRLFNVKTPSMEQAIGLLSGGNQQKVAIARGLMTRPKVLILDEPT  423

Query  422  RGVDVGAKREIYTLMNELTERGVAIIMVSSELPEILGMSDRIIVVHEGRISGEIHAREAT  481
            RGVDVGAK+EIY L+N+    G++II+VSSE+PE+LGMSDRIIV+HEG +SGE    +AT
Sbjct  424  RGVDVGAKKEIYQLINQFKADGLSIILVSSEMPEVLGMSDRIIVMHEGHLSGEFTREQAT  483

Query  482  QERIMTLATG  491
            QE +M  A G
Sbjct  484  QEVLMAAAVG  493
```